

**CENTRO ESTADUAL DE EDUCAÇÃO TECNOLÓGICA  
PAULA SOUZA**

**FACULDADE DE TECNOLOGIA DE INDAIATUBA**

**DR. ARCHIMEDES LAMMOGLIA**

**Caroline Raissa Ferraz**

**Alexandre**

**ANÁLISE DE DADOS EM R**

**Analisando as Notas do Exame**

**INDAIATUBA  
Setembro/2022**

**Caroline Raissa Ferraz**

**Alexandre**

## **Analisando as Notas do Exame**

Este trabalho visa  
como seu principal  
objetivo contribuir  
para a compilação de  
todos os recursos e  
ensinamentos  
vinculados a disciplina  
de Sistemas da  
Informação da FATEC  
Indaiatuba Dr.  
Archimedes  
Lammoglia.

**INDAIATUBA  
Setembro/2022**

## **INTRODUÇÃO**

A base de dados escolhida ajuda a entender o impacto de diferentes fatores, tais como o gênero do avaliado; etnia; nível de escolaridade dos pais; acesso a almoço e se havia feito um curso preparatório; nos índices de notas de um teste. A base possui cerca de 1000 conjuntos de dados cadastrados.

Ela foi escolhida por ser de fácil visualização e percepção para o estudo de Análise de Dados. No desenvolver deste trabalho apresentarei a análise técnica e resultados em gráficos e números dos dados coletados.

## **OBJETIVO**

O objetivo do estudo é aprender e compreender a ciência na qual transforma dados numéricos e qualitativos em uma conclusão para solução de problemas. A base de dados nos ajuda a entender colocando em prática esse tipo de análise

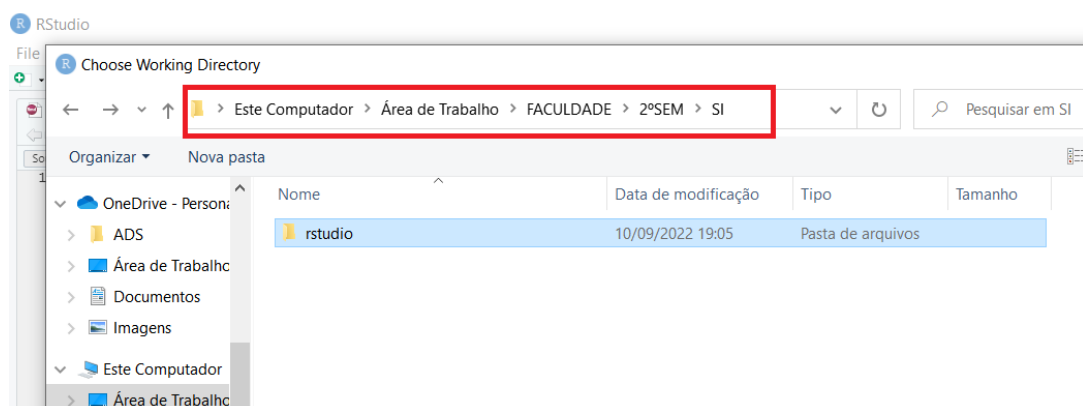
## METODOLOGIA

Primeiro foi estudado a base em linguagem R pela plataforma Udemy – Formação de Cientista de Dados.

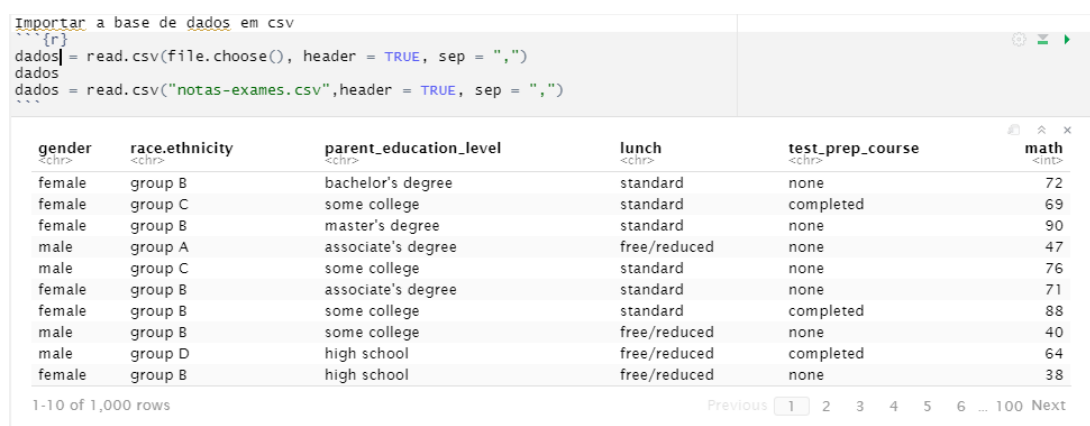
Com a base em linguagem R, foi escolhido uma base de dados para estudos.

Para fazer a importação dos dados, primeiramente foi mudado o diretório do trabalho para a pasta onde a base de dados se encontrava.

Session -> Set Working Directory -> Choose Working Directory



A base de dados foi importada para o software usando o código de importação para base de dados em texto como mostra a imagem a seguir:



Em primeiro momento os nomes das colunas foram trocados de inglês para português.

```

##{r}
#Dar os nomes corretos as colunas
colnames(dados) = c("genero", "raca/etnia", "Nível de educacao dos pais", "almoco", "curso preparatorio para testes", "matematica")
head(dados)

```

	genero <chr>	raca/etnia <chr>	Nível de educacao dos pais <chr>	almoco <chr>	curso preparatorio para testes <chr>
1	female	group B	bachelor's degree	standard	none
2	female	group C	some college	standard	completed
3	female	group B	master's degree	standard	none
4	male	group A	associate's degree	free/reduced	none
5	male	group C	some college	standard	none
6	female	group B	associate's degree	standard	none

6 rows | 1-6 of 6 columns

Após isso foi feito a análise dos dados, para verificar se havia algum erro na base, e assim caso necessário fosse feito a limpeza e o tratamento.

A verificação foi feita gerando gráficos de barra correlacionando cada coluna com quantidade de dados carregado no banco, para a verificação.

```

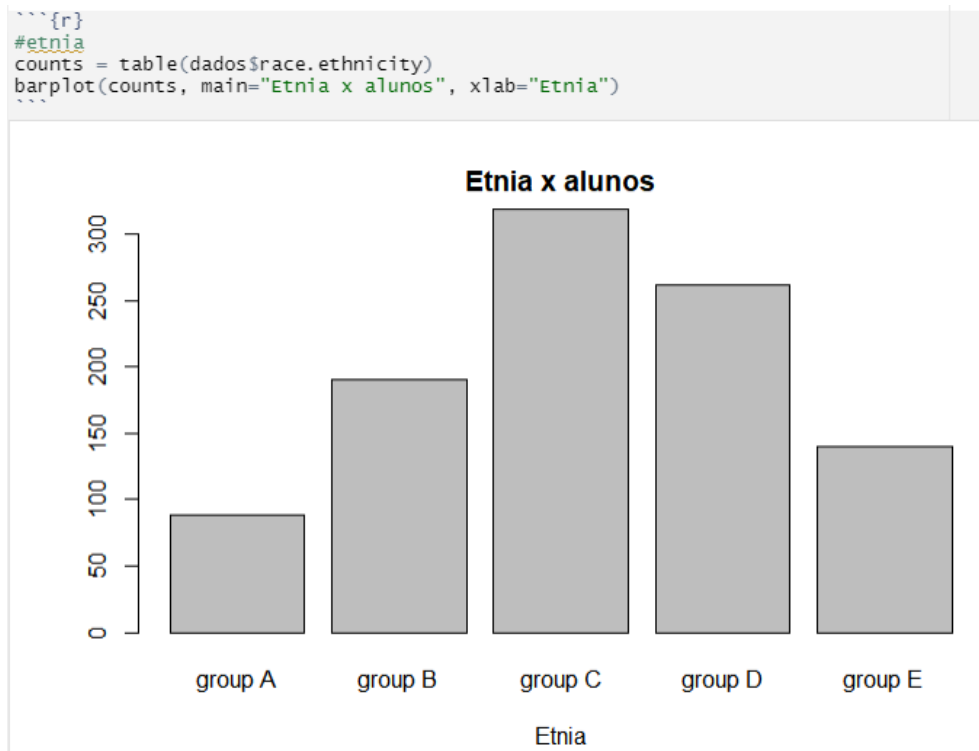
##{r}
#genero
counts = table(dados$gender)
barplot(counts, main="Gênero x alunos", xlab="gênero")

```



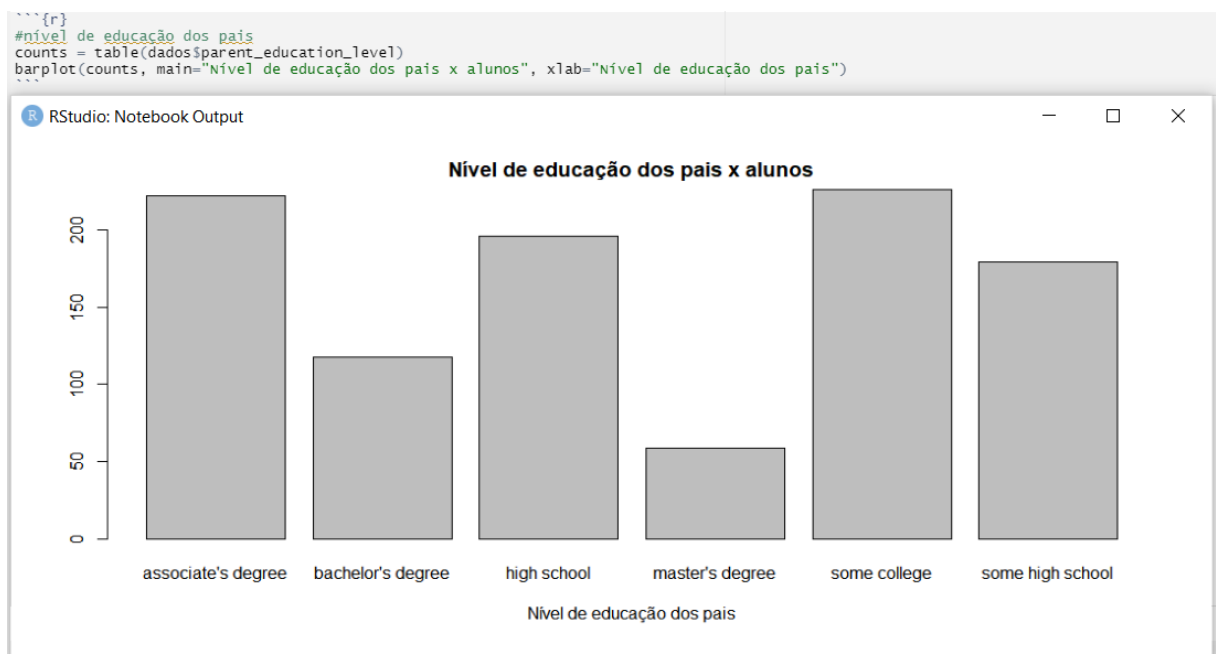
### Gráfico de barra 1 – Verificação de dados relacionais a gênero x alunos

O gráfico de verificação de dados relacionais a gênero x alunos é distribuído em duas categorias: feminino e masculino. O gráfico não mostra nenhuma divergência com os dados.



### Gráfico de barra 2 – Verificação de dados relacionais a etnia x alunos

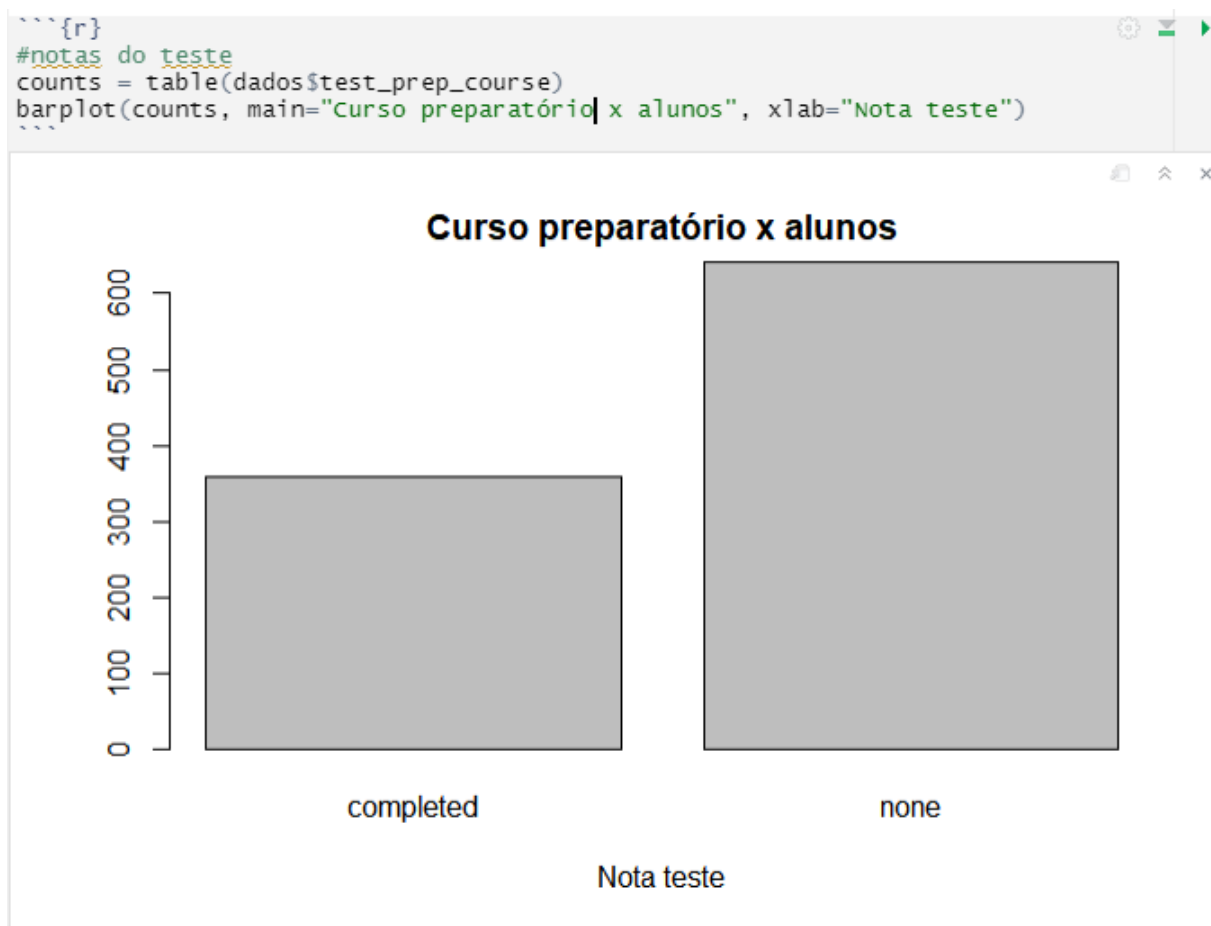
Houve a verificação da coluna etnia x alunos, onde os dados são distribuídos em grupos: Grupo A, Grupo B, Grupo C, Grupo D e Grupo E. Não há inconsistência nos dados.



### Gráfico de barra 3 – Verificação de dados relacionais nível de educação dos pais x alunos

Em relação a verificação de dados relacionais ao nível de educação dos pais x

alunos os alunos estão distribuídos dentro de seis subcategorias relacionadas ao nível de educação dos pais, são elas grau associado, bacharelado, ensino médio, mestrado, alguma faculdade e alguma escola de ensino médio. Nessa validação também não foi encontrado inconsistências nos dados seguindo para as próximas verificações.

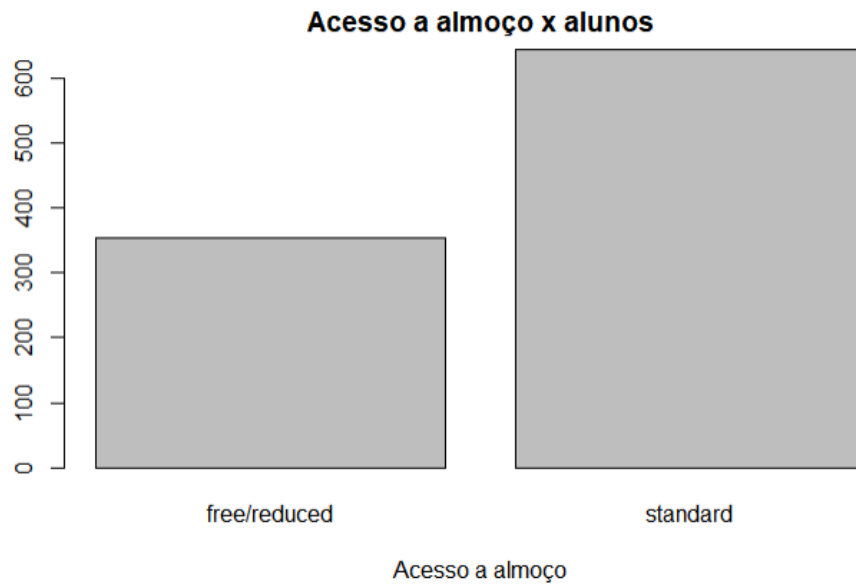


**Gráfico de barra 4 – Verificação de dados relacionais curso preparatório x alunos**

Na verificação de dados relacionais ao acesso a curso preparatório x alunos também não foi encontrado discrepâncias nos dados. Os alunos foram distribuídos em duas subcategorias, os que tiveram acesso ao curso (completed) e os que não tiveram (none).



```
##{r}
#se tinham acesso a almoço
counts = table(dados$lunch)
barplot(counts, main="Acesso a almoço x alunos", xlab="Acesso a almoço")
```



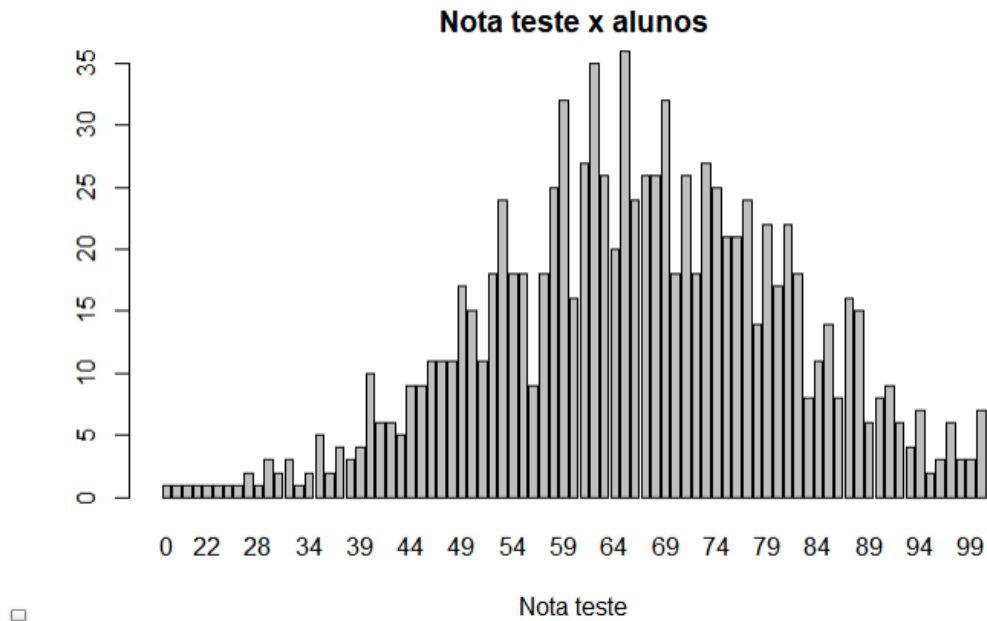
### Gráfico de barra 5 – Verificação de dados relacionais almoço x alunos

A verificação de dados relacionais ao nível de educação dos pais x alunos também não foi encontrado discrepâncias nos dados. Os alunos foram distribuídos em duas subcategorias, em alunos que recebiam almoço grátis ou com preço reduzido e em alunos que tinham almoço com preço padrão.

```

{r}
#notas do teste
counts = table(dados$math)
barplot(counts, main="Nota teste x alunos", xlab="Nota teste")

```



### Gráfico de barra 6 – Verificação de dados relacionais nota teste x alunos

A verificação de dados relacionais as notas de teste x alunos também não foi encontrado discrepâncias nos dados. Os alunos foram distribuídos sob categorias nos valores de 0 a 100, que seria o peso do teste.

Como os dados não apresentaram nenhuma inconsistência não houve a necessidade de fazer a limpeza e o tratamento.

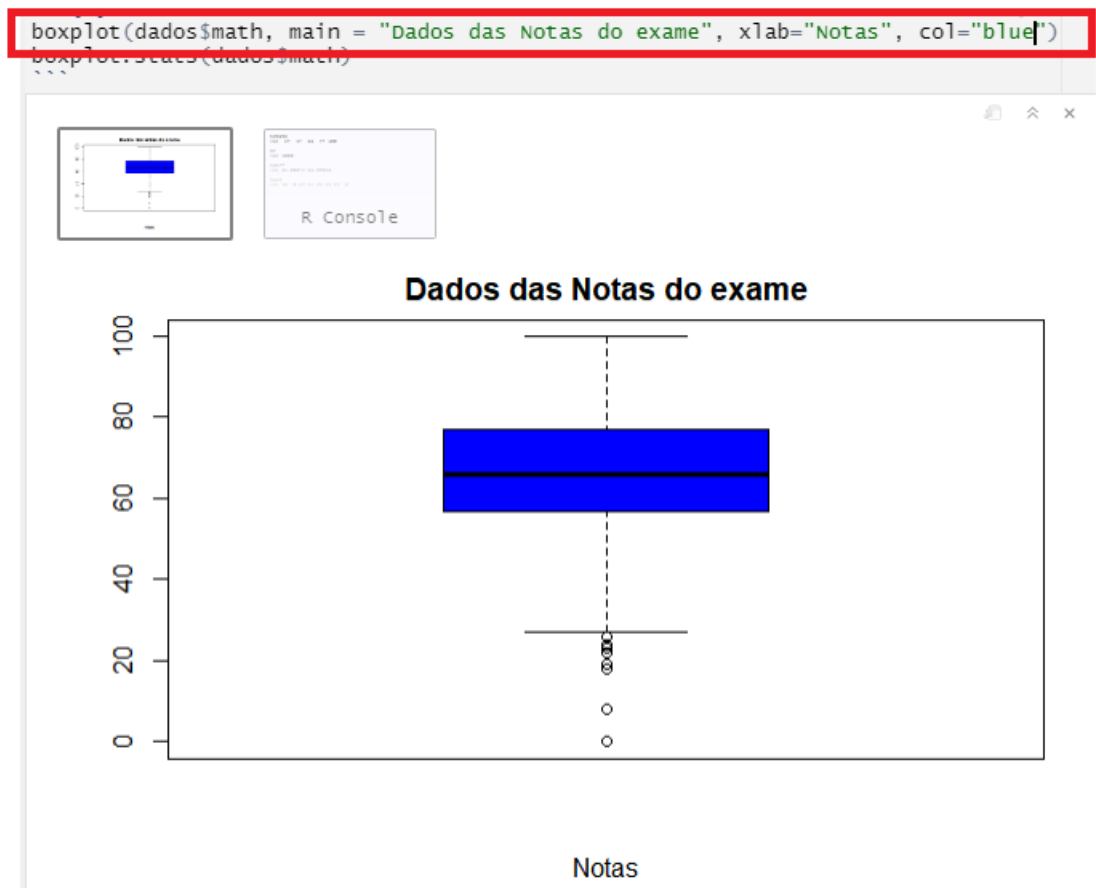
Após a verificação foi dado o início a etapa de escolha das funções que constituiriam construção da análise.

Para isso, foi levantado um conjunto de questões a serem respondidas durante a análise, são elas:

- ✓ Qual foi o resultado das notas? E quem são esses alunos?
- ✓ Qual fator pode ser considerado o que mais interfere as notas?
- ✓ Como os fatores interferem com as notas?
- ✓ Como pode ser definida a classificação etnia? Quem são os alunos distribuídos por essas classes?

Com o intuito de responder as questões levantadas foi gerado o resultado total das notas, de forma que se chegue à conclusão de como as notas estão distribuídas

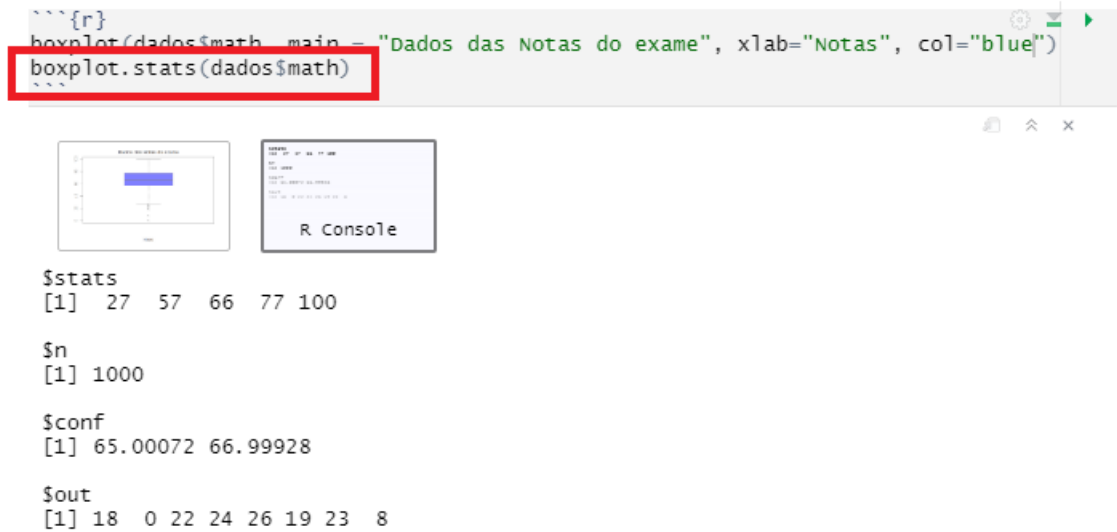
de acordo com a quantidade de alunos. Para obter uma visão geral das notas foi usado o gráfico boxplot (diagrama de caixa). A variável utilizada foi notas(**math**), **main** diz respeito ao título do diagrama e **xlab** ao nome da variável escolhida para análise, **col** a cor do diagrama.



### Diagrama de caixa 1 – Variação relativa das notas dos avaliados

De acordo com o diagrama valor máximo das notas chegou a 100 e o mínimo está entre 20 e 40. Com outliers abaixo do valor mínimo de nota chegando a 0. A mediana se aproxima de 60.

A média das notas se encontram entre os valores de 60 a 80 de nota.



### Dados do diagrama de caixa 1 – Variação relativa das notas dos avaliados

De forma mais clara, com a função *boxplot.stats* mostra os dados usados para gerar o diagrama.

*\$stats* : valores usados para a montagem do diagrama

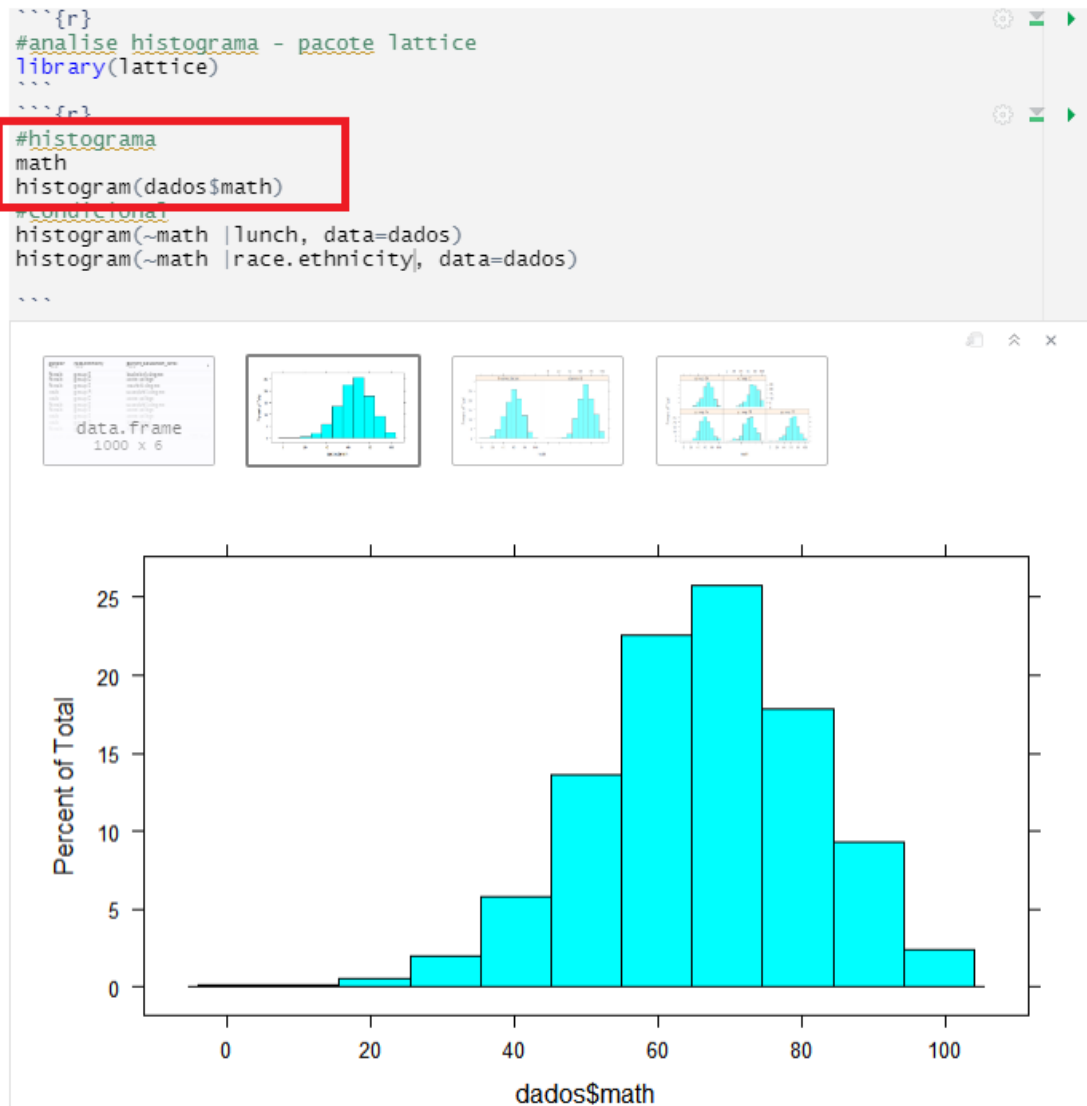
*\$n* : quantidade de dados

*\$conf* : mediana

*\$out* : outliers (dados que se diferenciam drasticamente dos outros)

Com os dados dispostos fica lucido a forma como eles são distribuídos na leitura. Dentro do diagrama a menor nota chegou a 27 e a maior 100, porém existem avaliados que tiraram notas menores (outliers), chegando a zerar no teste. A nota mediana é 66.

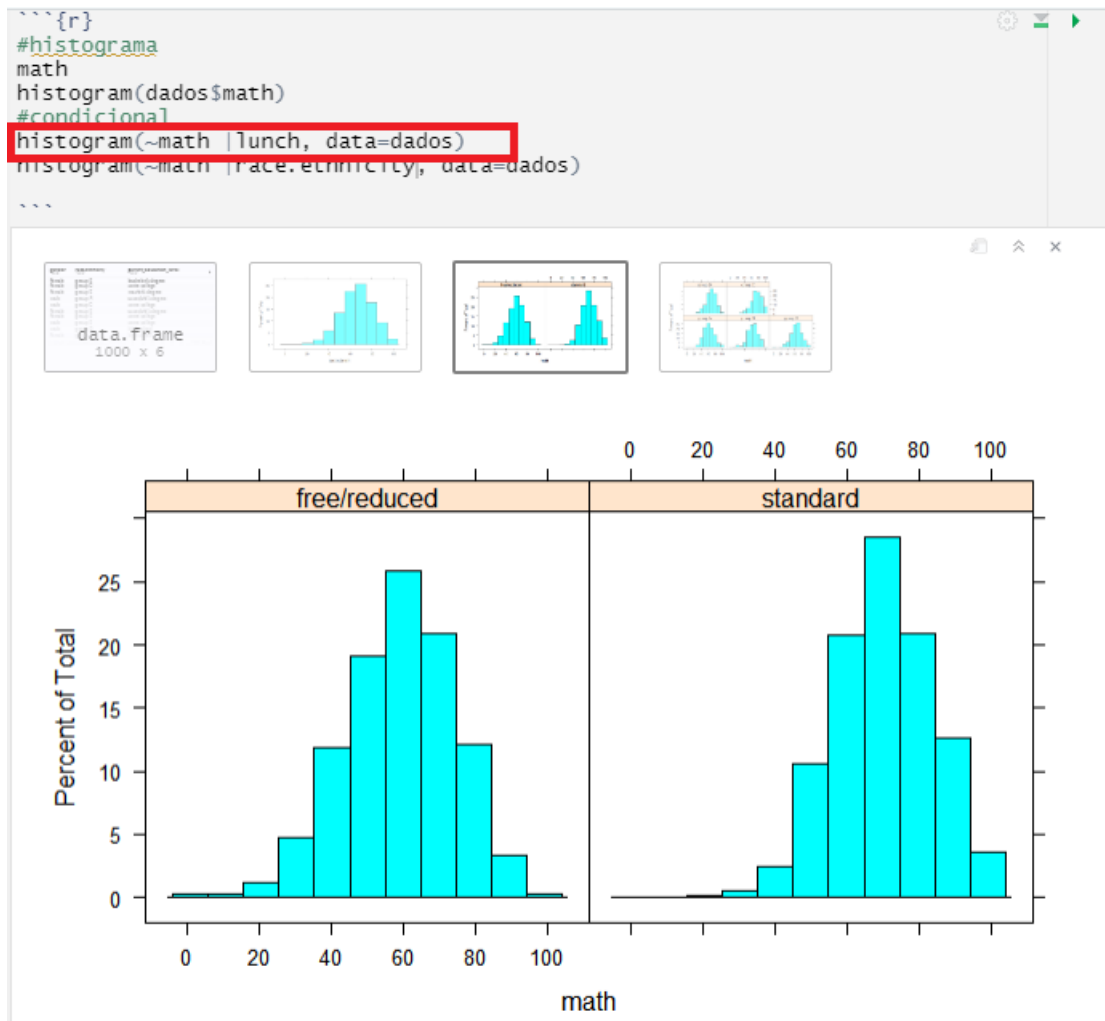
Para analisar o percentual das notas com as variáveis foi usado histograma.



### Histograma 1 – Percentual de notas

Esse histograma tem como o objetivo mostrar o percentual de notas alcançadas com os testes. Aqui se pode perceber que o maior percentual de alunos tirou notas entre 60 e 80 (25%), também chegamos à conclusão de que menos de 5% dos alunos chegaram a gabaritar e a zerar a prova.

Usando como base o histograma 1, foi gerado dois condicionais para análise mais aprofundada dos dados. Foi usado uma variável contínua que no conjunto de dados são as notas (**math**), só que a mesma quer ser analisada condicionalmente com o conjunto de dados referente a almoço (**lunch**), etnias (**race.ethnicity**), nível de educação dos pais (**parent\_education\_level**) e acesso a curso preparatório (**test\_prep\_course**), do conjunto de dados (**data = dados**)




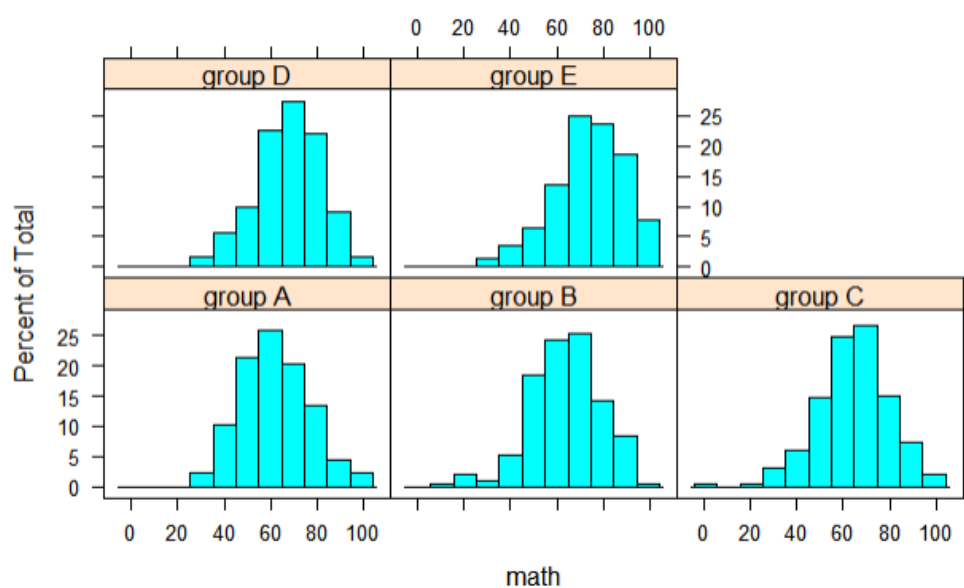
## Histograma 2 – Acesso ao almoço interfere no rendimento das notas

Esse histograma tem como o objetivo mostrar o percentual de notas alcançadas relacionando ao acesso dos alunos ao almoço.

Os alunos que tinham acesso a almoço grátis ou com preço reduzido tiveram em seu maior número nota 60 (mais ou menos 25% do total de alunos). Menos de 2% do total dos alunos conseguiram gabaritar a prova.

Os alunos que tinham acesso padrão ao almoço tiveram em seu maior percentual notas entre 60 e 80 (mais de 25% do total de alunos). Entre 3% e 5% dos alunos conseguiram gabaritar a prova.

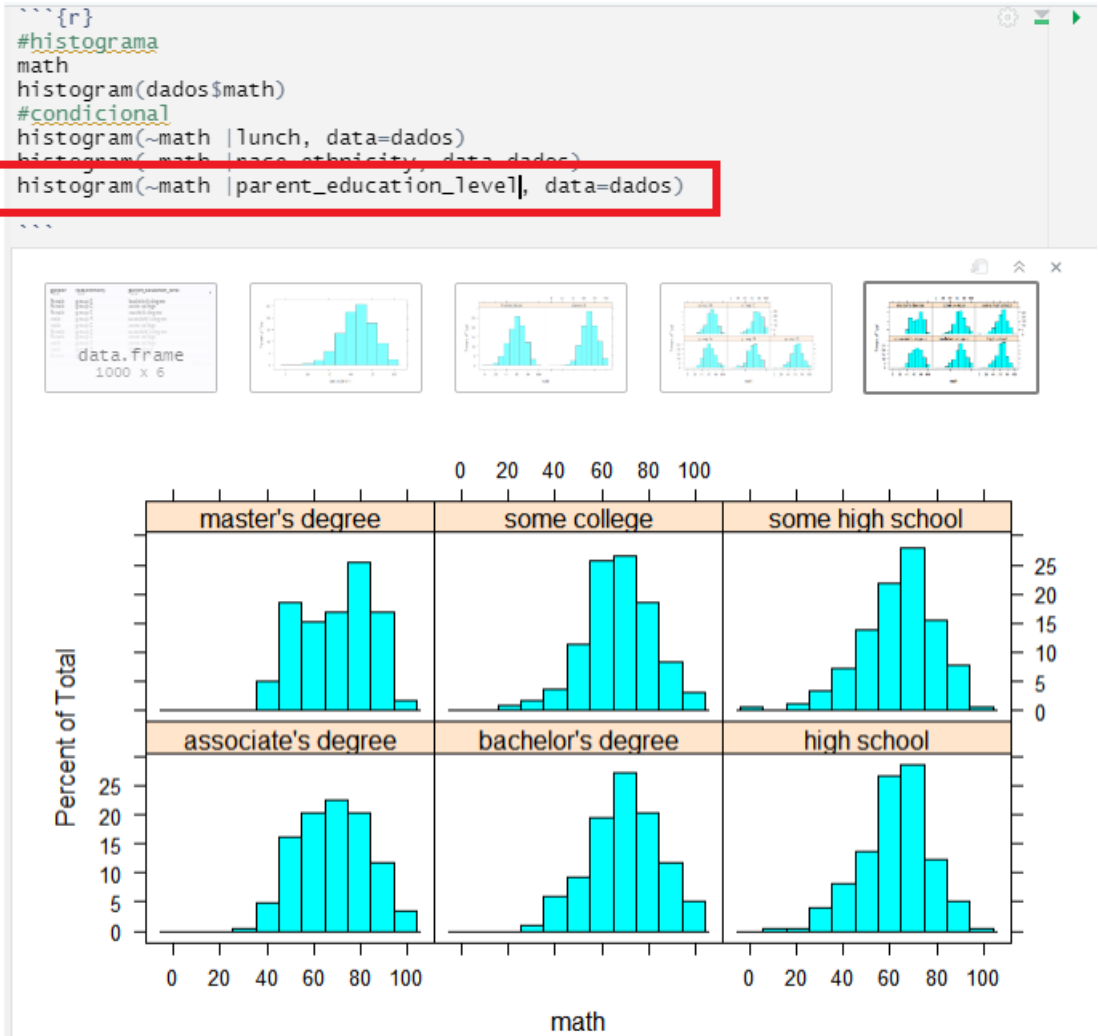
```
{r}
#histograma
math
histogram(dados$math)
#condicional
histogram(math | lunch, data=dados)
histogram(~math | race.ethnicity, data=dados)
```

### Histograma 3 – Interferência das classes de etnia em relação ao rendimento das notas

As etnias que possuíram um rendimento maior foram Grupo A, D e E, sendo o Grupo E o que obteve uma porcentagem maior de alunos que alcançaram nota maior que 80.

Os grupos B e C possuem um percentual maior entre notas de 40 e 80, os grupos também possuem percentual de alunos que tiraram notas abaixo de 20.



#### Histograma 4 – Dados relacionais as notas dos exames com o nível escolar dos pais

O percentual de maiores notas se encontra com os alunos que possuem pai com bacharelado (bachelor's degree), grau associado (associate's degree) e mestrado (master's degree). Os que possuem alguma faculdade e ensino médio completo o percentual total se distribuem entre as notas e se concentram na média de 60.

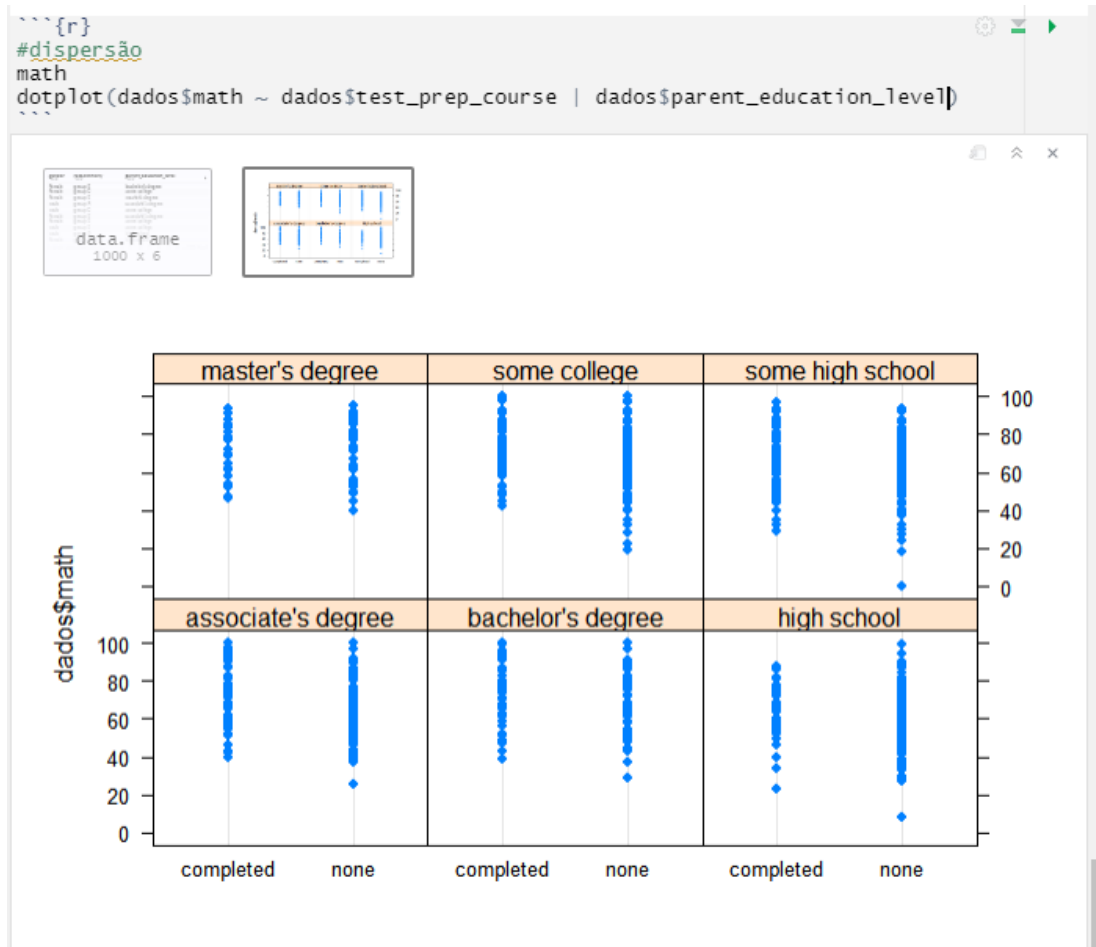




### Histograma 5 – Dados relacionais as notas dos exames com o acesso a curso preparatório

Os alunos que possuem o curso preparatório completo alcançaram em seu maior percentual (mais de 60% do total de alunos) notas acima de 60. Já os alunos que possuem curso preparatório incompleto as notas mais distribuídas (mais de 55% do total de alunos) estão entre as notas 40 e 80.

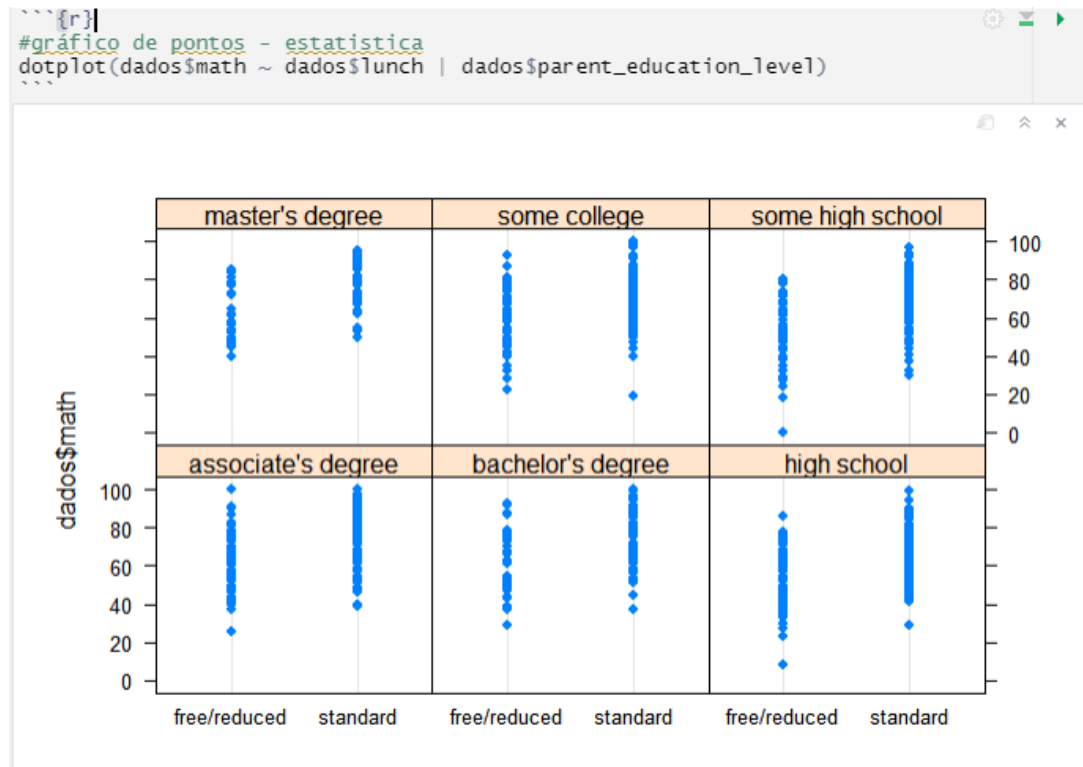
Para analisar as estatísticas dos dados foi usado o gráfico de pontos. Os valores (pontos) aparecerão nos valores em que o dado se encaixa. Foi usado a função `dotplot()`, usando as variáveis **`dados$math`** – dados das notas, **`dados$test_prep_course`** (dados dos cursos preparatórios), **`dados$parent_education_level`** (dados do nível de educação dos pais).



### Gráfico de pontos 1 – Acesso a ensino x notas

O gráfico mostra três variáveis, são elas: notas, ensino dos pais e se o aluno teve acesso a curso preparatório. Ao realizar a análise se pode perceber que a proporção de alunos sem curso preparatório é maior que a de alunos com curso preparatório completo. E que a proporção de alunos com pais que possuem somente ensino médio é maior do que comparado a avaliados que possuem pais com bacharelado e mestrado.

Em questão a análise conjunta dos dados o gráfico mostra que a maior proporção dos avaliados não possuem curso preparatório e seus pais possuem apenas o ensino médio e somente colégio e que a menor proporção dos avaliados possuem ensino preparatório completo e seus pais possuem mestrado completo.



### Gráfico de pontos 2 – Acesso a alimentação x notas

O gráfico mostra três variáveis, são elas: notas, ensino dos pais e o acesso a almoço. Ao analisar o gráfico se pode perceber que a maior parte dos avaliados possuem acesso padrão ao almoço e que eles se encontram em sua maioria em avaliados que possuem pais com colegial (some college), ensino médio (high school) e grau associado (associate's degree).

## RESULTADOS

Dentro da base de dados a maior nota dos avaliados chegou a 100 e a menor nota chegou a 0 (mesmo que não haja uma quantidade considerável de avaliados com essa nota), a mediana das notas foi 66 e a média ficou entre 60 e 80.

Em geral a maior proporção dos avaliados não possuem curso preparatório e seus pais possuem apenas o ensino médio completo, o que pode se associar a média de notas tiradas entre 60 e 80. Em sua menor proporção os avaliados possuem ensino preparatório completo e pais com mestrado completo, considerando que essa proporção chegou a tirar notas maiores que 80.

Os grupos A e E possuem notas maiores, podendo possuir padrões de vida melhores, pois sua proporção total são os grupos que possuem menos integrantes. O grupo C é onde se concentra a maior parcela dos avaliados, eles obtiveram mais variações de notas, podendo considerar que os integrantes possuem um estilo de vida padrão.

As maiores notas foram adquiridas por aqueles que possuem pais com bacharelado (bachelor's degree), grau associado (associate's degree) e mestrado (master's degree).

## **CONCLUSÃO**

Em suma, os avaliados em sua maioria possuíam estilo de vida padrão, pais com ensino médio e bacharelado completo, acesso padrão a almoço e não tinham acesso a curso preparatório, chegando a obter resultados de exame entre 60 e 80.

Podemos concluir que fatores econômicos podem interferir no rendimento das notas finais, já que os avaliados que foram considerados com condições financeiras maiores foram os que conseguiram alcançar as maiores notas (acima de 80).

A Análise de dados é uma ferramenta de trabalho muito eficaz para obter resultados e fazer previsões. Dentro dessa base de dados, além dos resultados obtidos, podemos considerar ações para melhorar as notas para os próximos exames. Um exemplo de ação a ser considerada é em relação a curso preparatório, se todos os avaliados tivessem acesso, a densidade de notas poderia se concentrar para acima de 80, já que essa variável se mostrou interferente com o resultado.

## REFERÊNCIA TEÓRICA

**“Analisando as notas do exame”**. Disponível em: [Analisando as notas do exame | Kaggle](#). Acessado em 07 de setembro de 2022.

Amaral, Fernando. **“Formação de Cientista de Dados: O curso completo [2022]”**. Disponível em: [Formação Cientista de Dados: O Curso Completo \[2022\] | Udemy](#). Acessado em 07 de setembro de 2022.

**Linguagem R: entenda como funciona e principais aplicações**. Disponível em: [Linguagem R: entenda como funciona e principais aplicações - Remessa Online](#). Acessado em de setembro de 2022.